

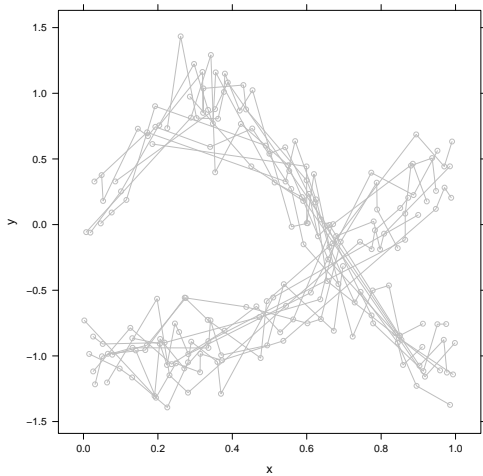
Simplified Clustering with Dirichlet Process and Other Process Mixtures

Matt Shotwell

Medical University of South Carolina
Biostatistics and Epidemiology

November 30, 2009

Problem



Dirichlet process mixture model (DPM)

$$\begin{aligned} \mathbf{y}_i &\sim N(\Omega(\mathbf{x}_i)\beta_i, \tau_i) \\ (\beta_i, \tau_i) &\sim G \\ G &\sim DP(G_0) \end{aligned}$$

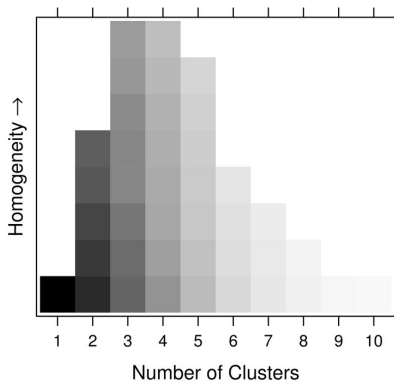
Product partition model (PPM)

$$\begin{aligned} \mathbf{y}_i | z_i = k &\sim N(\Omega(\mathbf{x}_i)\beta_k, \tau_k) \\ (\beta_k, \tau_k) &\sim G_0 \\ P(\mathbf{z}) &\propto \prod_{k=1}^r c(\mathbf{z}) \end{aligned}$$

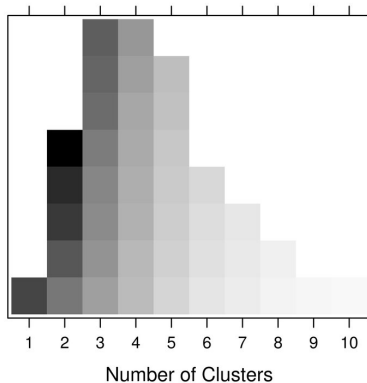
$$c(\mathbf{z}) = \begin{cases} (n_k - 1)!/r! & \text{Dirichlet} \\ n_k/r! & \text{Cluster} \end{cases}$$

Alternate Cohesions $c(\mathbf{z})$

'Dirichlet' Prior



'Cluster' Prior



$$P(\mathbf{z}|\mathbf{y}) \propto \prod_{k=1}^r c(\mathbf{z})P(\mathbf{y}_{(k)}|\mathbf{z})$$
$$P(\mathbf{y}_{(k)}|\mathbf{z}) = \iint \prod_{k=1}^r P(\mathbf{y}_{(k)}|\beta_k, \tau_k)dG_0$$

Profile inference

$$\hat{\mathbf{z}} = \arg \max P(\mathbf{z}|\mathbf{y})$$
$$P(\beta_k, \tau_k|\mathbf{y}, \hat{\mathbf{z}}) \propto \prod_{k=1}^r P(\mathbf{y}_{(k)}|\beta_k, \tau_k)P(\beta_k, \tau_k)$$

Yeast Cell Cycle Data - Spellman *et al.* (1998)

